IDENTIFICATION OF ARABIC DIALECTS AND EXPERIMENTAL DETERMINATION OF DISTINCTIVE CUES

Melissa Barkat
Laboratoire Dynamique Du Langage (France)

ABSTRACT

This paper deals with the identification of Arabic vernaculars based on analysis of natural speech. Six Arabic dialects were chosen from two major dialect areas of the Arab World (i.e. Maghreb vs. Middle-East). Speech data were obtained by recording spontaneously produced narration (i.e. description of pictures) by 12 Arabic speakers from 6 different countries. Samples of speech (complete utterances) were then selected so as to make up the stimulus material for a perception test. The results of the experiment show that native Arabic speakers are successful in identifying the dialectal varieties used on the basis of cues of different kinds (segmental, syntactic, lexical and prosodic). This perceptual experiment must be regarded as a first step towards the determination of a set of reliable cues for the Automatic Identification of Arabic Dialects.

1. INTRODUCTION

Most automatic recognition systems in existence (for French, English…) are based on linguistic forms selected on standard varieties of these languages But dialectal varieties of these languages can, formally speaking be quite distant from the standard language.

Because dialects can be structurally very distant from the norms of Modern Standard Arabic (MSA), it is necessary to develop methods for the Automatic Identification of these dialects as linguistic systems separate from MSA. The aim of the present paper is to demonstrate that Arabic speakers from different part of the Maghreb and the Middle-East are able to identify very accurately the geographical origin of a given Arabic speaker even though the degree of mutual intelligibility between the two speakers is very low or nil. The ultimate objective of the proposed research is to identify a set of salient cues relevant for the automatic identification of Arabic dialects.

2. DIALECT GEOGRAPHY OF THE ARAB WORLD

All Arabic dialects show some degree of innovation and / or conservatism when compared with MSA. These can include phonetic, syntactic, lexical and prosodic features. In addition to the major dialect groups mentioned above, one could also distinguish between Bedouin dialects and those spoken by co-territorial sedentaries [2]. These distinctions makes it possible for the 'native ear' to identify the dialectal variety used by a given speaker.

3. LINGUISTIC CORPUS AND METHODOLOGY

In order to represent the two geographical areas mentioned above, the six dialectal varieties of Arabic mentioned in table 1 have been chosen. An acoustic database was elaborated by recording 12 speakers (male and female) who were required to produced a short narration by describing, in their native dialect, a book made of 15 pictures[1].

Western Area	Eastern Area
Morocco	Syria
Algeria	Lebanon
Tunisia	Jordan

Table 1. Dialectal varieties represented (for the stimuli and the subjects).

Recordings were digitized at 22KHz, 16 bits, mono. 96 samples of speech (i.e. complete utterances from 5 to 30 seconds) were extracted and presented at random as stimuli for a perceptual experiment to 18 subjects, native speakers of Arabic from the same six countries. The subjects' tasks entailed three stages :

(i) to identify the stimuli in terms of geographical area (western vs. eastern)
(ii) to associate the stimuli in term with specific country as specified in table 1.
(iii) to define (when possible) the segmental, prosodic and / or lexical cues that allow the subject to identify the dialect in question.

In addition to determining the distinctive acoustic cues relevant for the identification of these Arabic dialects, we intended to verify the following theoretical issues :

(i) be able to identify the relevant On being exposed to a stimuli, the subjects would dialect area.
(ii) Within an area, the best results would be obtained for dialects closest to the subject's native variety of Arabic.
(iii) Incorrect identifications would occur exclusively in relation to dialects that are geographically close.

4. RESULTS

4.1. Dialectal identification

4.1.1. Identification per area : We observe 97 % of the western stimuli and 99% of the middle-eastern ones were identified correctly by the subjects: These results confirm the assumption that the linguistic classification of the Arabic dialect area adopted for the study is significant for native speakers on the basis of phonetic an other cues.

4.1.2. Regional identification : The following figures (2a to 2f) show the rates of correct identification per country obtained by the 18 subjects according to their origin.

Our results, although asymmetrical from one area to the other, reveal that the best rates are obtained :

Fig.2(a) Moroccan Stimuli

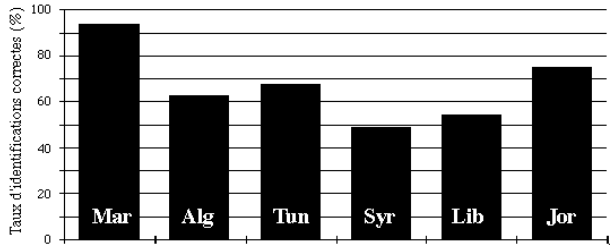


Fig.2(b) Algerian Stimuli

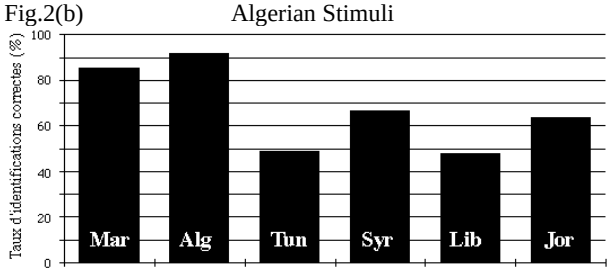


Fig.2(c) Tunisian Stimuli

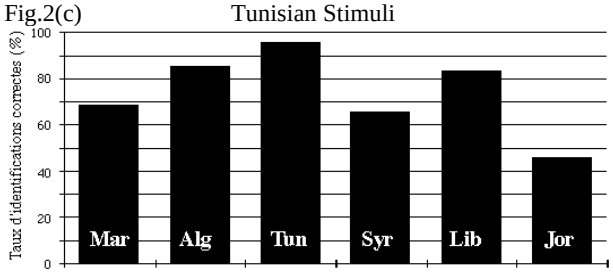


Fig.2(d) Syrian Stimuli

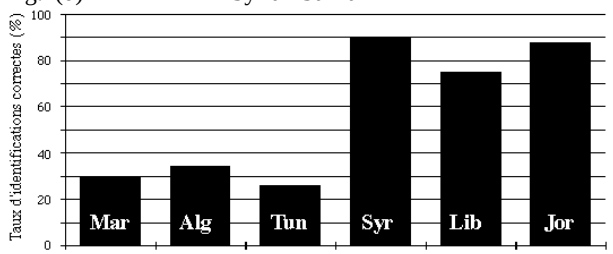


Fig.2(e) Lebanese Stimuli

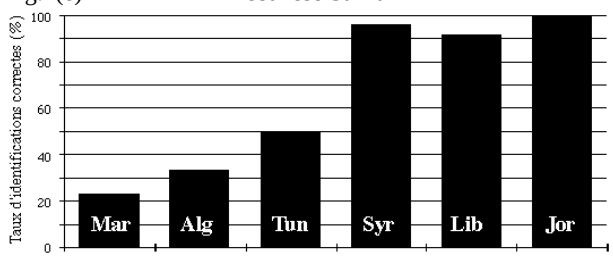
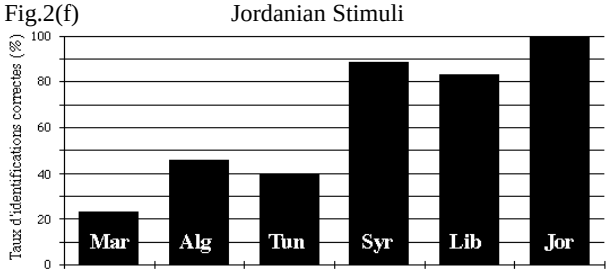


Fig.2(f) Jordanian Stimuli



(i) for the identification of their own dialectal variety
(ii) for the dialects belonging to the same area than their own.

Nevertheless, it is interesting to note that middle-eastern subjects' results as far as regional particularism is concerned are homogenous (i.e. good rates for the recognition of western and middle-eastern dialects were obtained) whereas speakers of maghrebine Arabic achieved lower rates of correct identification. This could be explained by the use of Arabic as their unique medium of communication in the Arab East as opposed to the use of foreign languages like French or English by Maghrebian speakers communicating with middle-eastern speakers. In practice, this entails only limited familiarity with the peculiarities of Eastern Arabic.

4.2. Error analysis

Wrong identifications occurred in relation to dialects from the same geographical area (i.e. Moroccan and Algerian for the western area ; Lebanese and Syrian for the eastern one). The incorrect identifications are due to the fact that in these areas dialectal boundaries do not correspond exactly to national borders [2].

Indeed, dialectal pairs like Moroccan and Algerian or Lebanese and Syrian share several common features (like for example the presence of a glottal stop [ʔ] in Lebanese and Syrian corresponding to the MSA uvular stop [ق] and to the Jordanian [g] ; or the centralization of the maghrebine vocalic systems typical of Algerian and Moroccan dialects [3].

4.3. Experimental determination of distinctive cues

The dialect spoken by the subjects constitutes the auditory filter through which they identified the features of the other varieties. These features can be a priori considered as reliable for the automatic identification of Arabic dialects in so far as they have been perceived by Arabic speakers as relevant for the distinction of the six different varieties under study. It is nevertheless, important to consider these features as typical for our informants' dialectal varieties and not as generally typical for an entire country.

4.3.1. Phonetic cues mentioned by the subjects

• 3 different realizations of the MSA uvular stop [ق] according to dialect.

Cue	Countries	Examples & translation
[ق]	Algeria, Morocco, Tunisia	[قرون] "horns"
[ق] >[g]	Jordan	[grun] "horns"
[ق] > [ʔ]	Syria, Lebanon	[ʔrun] "horns"

• Different treatment of the MSA labiodental fricatives [θ,ð,ظ]. Either conservation/change of the labiodental fricatives or fusion with the corresponding stops [θ,ð,ظ] > [t,d,dˤ].

Cues	Countries	Examples & translation
[θ]	Algeria, Jordan	[θamma] "then"
[θ]>[t]	Morocco, Lebanon, Syria	[tamma] "then"
[θ]>[f]	Tunisia	[famma] "then"
[ð]	Algeria, Jordan	[ðiːb] "wolf"
[ð]>[d]	Lebanon, Morocco	[diːb] "wolf"
[ðˤ]	Algeria, Jordan	[ðˤufðˤaʕ] "frog"
[ðˤ]>[dˤ]	Lebanon, Morocco, Syria	[dˤufdˤaʕ] "frog"

• Affrication of the voiceless dental stop [t] > [ts]

Cue	Countries	Examples & translation
[t] > [ts]	Algeria	[tsksar] "it broke"

• Affricated or velar articulation Arabic [ʒ] [9].

Cue	Countries	Examples & translation
[g]	Morocco	[gʕas] "he sat"
[ʧ]	Syria	[waʧʕa] "his face"
[dʒ]	Algeria	[dʒʕas] "he sat"

• The reflex [zˤ] in the lexical item [sˤɣiːr]

Cue	Countries	Examples & translation
[sˤ] > [zˤ]	Syria	[sˤɣiːr] > [zˤɣiːr] "small"

• Nasalization of the lateral approximant [l] > [n]

Cue	Countries	Examples & translation
[l] > [n]	Lebanon	[mliːha] > [mniːha] "good (feminine)"

•Centralization of short vowels in CVC structures [4]

Cue	Countries	Examples & translation
[i, a, u] > [ə]	Algeria, Morocco and Tunisia	[sakat] > [skat] "he was silent"

• Fronting of the back vowel [a] in final and median position (i.e. "imala" process) [3]

Cue	Countries	Examples & translation
[a] > [e , ɛ]	Middle-East and Tunisia (in monosyllabic words only)	[kbiːre] "tall (feminine)" [lmɛː] "water"

• Shift from [aj]> [eː]

Cue	Countries	Examples & translation
[aj] > [eː]	Lebanon	[beːt] "house"

4.2.3. Morpho-syntactic cues mentioned by the subjects

• Change in the quality of thematic vowels

Cue	Countries	Examples & translation
[a] > [i]	Lebanon	[harab] > [hirib] "he escaped"

• Imperfective marker

Cue	Countries	Examples & translation
[ki]	Morocco	[kiʕajjaʕ] "he is shouting"
[bi]	Lebanon	[bijirkod] "he is running"
∅	Algeria, Tunisia	[ʕajjaʕ] "he is shouting"

• Diminutive nominal scheme (insertion of the infix [ijjə])

Cue	Countries	Examples & translation
[ijjə]	Algeria	[klijjab] "puppy"

4.2.4. Lexical cues

Among the several isolated lexical items related by the subjects to a particular dialect, we selected the most useful for the development of a model of automatic recognition based on frequent word spotting.

• Possessive particle

Cue	Countries	Examples & translation
[dieːlo]	Morocco	"his"
[bteːʕo]	Lebanon, Syria, Jordan	"his"
[ntseːʕo]	Algeria	"his"
[ntɛːʕo]	Tunisia	"his"

4.2.5. Prosodic Cues

• Emphasis

Cue		Countries
Spread of emphasis	the whole word	Jordan
Spread of emphasis	Limited to a few segments in the word	Tunisia

• Syllable structure

The prosodic patterns in Western Arabic dialects are different from Middle-Eastern ones since the preferential syllabic structure is not the same (i.e. CVCVC in Middle-Eastern dialects vs CCVC in Maghrebian dialects). These different syllabic structures lead, on the one hand, to phonetic variations (i.e. centralization of short vowels in open syllables in Western varieties for example [4] and typical consonantal clusters), and on the other hand to different accentual treatments (on the first heavy syllable (i.e. CVV or CVC(C)) and up to the third mora from right to left in Eastern Arabic, on the first heavy syllable and never further than to the second mora in Western varieties) [5].

5. CONCLUSION

This study on the perceived areal diagnostic traits for vernacular Arabic highlighted the following points :

Speakers of Arabic are aware of these dialectal differences and generally succeed in locating their areal origins.

High scores are obtained for the identification of the varieties close to their native dialect.

Nevertheless, some dialectal frontiers can only be vaguely identified by country and are associated by speakers with more than one country.

In this study we have tried to focus attention on more narrowly diagnostic features. In the course of our research we shall attempt to evaluate the diagnostic potential of each feature and to identify what appear to us to be the most significant and useful for the automatic identification of Arabic dialects [6].

NOTES

[1]Mayer, M., (1980), « Frog, where are you ? », Dial Book for Young Readers, New-York.

REFERENCES

[1] Cohen D. (1968), "Langue Arabe" in Encyclopédie Universalis (T.II) pp. 201, Paris :

[2]Marçais Ph. (1977), “Esquisse grammaticale de l'arabe maghrébin”, Maisonneuve, Paris :

[3] Barkat M., Hombert, J.M & Taine-Cheikh C., (1997), “Détermination d'indices acoustiques robustes pour l'identification des parlers arabes : les variations phonétiques de la voyelle [a]”, in Actes des Journées d'Études Linguistiques, pp.141-146, Nantes :

[4] Barkat, M., (1999), “La dispersion vocalique comme indice discriminant pour l’identification automatique des parlers arabes”, in Actes des Premières Journées d’Identification de la Parole, Lyon (in press).

[5]Bohas, G. & Khoulougli, D.E., (1981), “Processus accentuels en arabe”, Analyses et théorie (1), Etudes arabes, Paris VIII, pp. 1-59

[6] Pellegrino, F., & Barkat, M., (1999) “Investigating dialectal differences via vowel system modeling : application to Arabic”, in proceedings of ICPhS99, San Francisco.